

U.S. DEPARTMENT OF COMMERCE
National Technical Information Service

PB-258 119

Analysis and Trial Application of
Correlation Methodologies for Predicting
Toxicity of Organic Chemicals

Franklin Inst Research Labs, Philadelphia, Pa

Prepared for

Environmental Protection Agency, Washington, D C

May 76

NTIS

Handbook on Aerosols
TID-26-608/PAT 141 p PC\$6.00/MF\$3.00

**A Preliminary Forecast of Energy Consumption
Through 1985**
PB-251 445/PAT 69 p PC\$4.50/MF\$3.00

Thank you for your interest in NTIS. We appreciate your order.

National Technical Information Service
U.S. DEPARTMENT OF COMMERCE
Springfield, Va. 22161
(703) 557-4650 TELEX 89-9405

Sub Total	
Additional Charge	
Enter Grand Total	

PB 258 119

STUDIES AND DATA APPLICATION IN CONDUCTING
RESEARCH FOR PREVENTING HAZARD OF
ORGANIC CHEMICALS

by EPA

Final Report

Prepared for

Office of Toxic Substances
Environmental Protection Agency
401 M Street
Washington, D. C. 20460

Contract No. 68-01-000

REPRODUCED BY
**NATIONAL TECHNICAL
INFORMATION SERVICE**
U. S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161



TECHNICAL REPORT DATA (Please read Instructions on the reverse before completing)		
1. REPORT NO. EPA-560/1-76-006	2.	3. RECIPIENT'S ACCESSION NO.
4. TITLE AND SUBTITLE Analysis and Trial Application of Correlation Methodologies for Predicting Toxicity of Organic Chemicals	5. REPORT DATE May, 1976 - Approved date	6. PERFORMING ORGANIZATION CODE
	8. PERFORMING ORGANIZATION REPORT NO. F-C3947	
7. AUTHOR(S) Paul N. Craig and John H. Waite	9. PERFORMING ORGANIZATION NAME AND ADDRESS The Franklin Institute Research Laboratories Science Information Services Department 20th & The Benjamin Franklin Parkway Philadelphia, Pennsylvania 19103	10. PROGRAM ELEMENT NO. 2LA328
12. SPONSORING AGENCY NAME AND ADDRESS Office of Toxic Substances Environmental Protection Agency 401 M Street, S.W. Washington, D. C. 20460	11. CONTRACT/GRANT NO. 68-01-2657	
	13. TYPE OF REPORT AND PERIOD COVERED Final	
		14. SPONSORING AGENCY CODE
15. SUPPLEMENTARY NOTES		
16. ABSTRACT An index to the literature on structure-activity correlation methods was prepared and is available through NTIS (PB 240-658). A study of each of the major methods was made to determine requirements for application to toxicity data. Simultaneously a study was made of available toxicity data and of physical-chemical properties shown to be useful in correlation studies. These evaluations suggested that the structural fragments contained in chemical structures should be considered in structure-activity relationship studies as well as the n-octanol partition coefficients. The U.S. Army C.I.D.S. computer-assigned fragment codes were utilized for this purpose. A prototype toxicity data base was selected from the 1974 Toxic Substances list for 687 compounds for which oral LD ₅₀ values were reported in the rat or mouse. The use of discriminant and multiple regression analyses following preliminary clustering gave useful results, but a new extension of the method called "substructural analysis" was used to predict the LD ₅₀ values in the rat for 21 of 23 test compounds within ± 1 log unit out of a range of 5 log units. This method can readily be adapted to computer operation, and is recommended for extension to other sets of toxicity data. Independent study of the same data by discriminant analysis is also recommended.		
PRICES SUBJECT TO CHANGE		
17. KEY WORDS AND DOCUMENT ANALYSIS		
a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Regression Analysis Discriminant Analysis Clustering Correlations Toxicity Pattern Recognition	Structure-Activity Relationships Substructural Analysis Chemical Structure Codes Partition Coefficients	06/20 06/04 12/01
18. DISTRIBUTION STATEMENT Release Unlimited	19. SECURITY CLASS (This Report)	21. NO. OF PAGES 52
	20. SECURITY CLASS (This page)	22. PRICE \$4.50 - \$3.00

EPA-560/1-76-006

ANALYSIS AND TRIAL APPLICATION OF CORRELATION
METHODOLOGIES FOR PREDICTING TOXICITY OF
ORGANIC CHEMICALS

Contract No. 68-01-2657

Project Officer
Frank D. Kover

Prepared for

Office of Toxic Substances
Environmental Protection Agency
Washington, D.C. 20460

May 1976

ia

NOTICE

This report has been reviewed by the Office of Toxic Substances, EPA, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Environmental Protection Agency, nor does mention of trade names for commercial products constitute endorsement or recommendation for use.

EXECUTIVE SUMMARY

This project was carried out in response to the needs of OTS for an analysis of the potential role which may be played by applications of correlation methodologies to a study of property-effect data in meeting their early warning objectives.

First, a literature survey was conducted to describe the state-of-the-art for the major methods of structure-activity correlation and of their potential utility for OTS needs; *second*, a study of available effects data to see whether their nature permits application of the various existing methods for correlation; *third*, a study of available physical-chemical properties and chemical structure fragment codes; and *fourth*, design, implementation and study of a prototype toxicity data base to test several methods side-by-side.

The literature survey resulted in an index of the literature on structure-activity correlation, which has been made available for distribution by the NTIS.

The study of available effects data has resulted in a rational approach to the organization and classification of the types of raw toxicity data encountered in this field. Such data do not often exist in sufficient quantities to permit application of classical statistical methods for analysis.

The study of many possible physical-chemical properties resulted in only one property which has been shown to provide almost universal applicability to structure-activity correlation problems, regardless of chemical structure — namely, the partition coefficient. A limited study of chemical structure-fragment codes showed that many could serve this function. The U.S. Army CIDS structure codes were chosen for the prototype data base because of their ready availability to EPA; other

systems, such as the Chemical Abstracts Services substructure fragments would have equal utility.

A prototype toxicity data base was developed using toxicity data for oral LD₅₀ values in the rat for 687 compounds selected from the NIOSH 1974 Toxic Substances List. Partition coefficients, molecular weights, and several hundred chemical structural fragment keys (obtained from the U.S. Army CIDS file at Edgewood Arsenal) were studied by two approaches to see whether meaningful predictions of toxicity could be made. The more successful was a modification of a method known as Substructural Analysis, first reported in 1974. Using the modified approach, rat toxicities for 21 out of 23 compounds studied were predicted with a reasonable degree of accuracy. Another group of 25 pairs of compounds, each pair containing one structural feature in common and representing the highest and lowest toxicity for that feature, was successfully ranked by this method.

The same data base was studied by readily available conventional methods of cluster analysis, discriminant analysis and regression analysis. Of these methods, discriminant analysis was found to permit separation of the small number of highly toxic substances from the bulk of the compounds, with a few false positives.

Recommendations are made for continuing the development and analysis of data in the prototype toxicity data bank. The simultaneous application of the modified substructural analysis method and discriminant analysis is recommended in extending this project to other types of toxicity, with emphasis on carcinogenesis and aquatic organisms toxicity data. As the number of sets of different toxic data in this data bank increases, the prediction of the degree of several kinds of toxicity for a novel substance will provide OTS with an early warning capability of increasing usefulness.

CONTENTS

<i>Section</i>	<i>Title</i>	<i>Page</i>
	EXECUTIVE SUMMARY.	iii
1	OBJECTIVES OF CONTRACT	1-1
	1.1 Background	1-1
2	WORK PERFORMED	2-1
	2.1 Literature Search	2-1
	2.2 Study of Available Toxicity Data.	2-1
	2.2.1 Toxicity Data in Registry of Toxic Effects of Chemical Substances (RTECS)	2-2
	2.2.2 Fish Muscle Accumulation Studies	2-3
	2.2.3 Carcinogenesis Data	2-3
	2.2.4 Mutagenesis Data	2-4
	2.2.5 Water Pollution Data	2-4
	2.2.6 Air Pollution Data	2-4
	2.2.7 Teratogenesis Data	2-5
	2.3 Study of Available Physical-Chemical Data	2-5
	2.3.1 Partition Coefficients	2-5
	2.3.2 Hammett Sigma Constants	2-6
	2.3.3 Taft Steric Constants.	2-7
	2.3.4 Molar Refractivity Values.	2-7
	2.3.5 Other Types of Physical Constants.	2-7
	2.4 Methods for Representing Chemical Structures.	2-8
	2.5 Evaluation of Methods for Structure-Activity Correlation	2-9
	2.5.1 Multiple Parameter Regression Analysis (Hansch Method)	2-9

CONTENTS (CONT.)

<i>Section</i>	<i>Title</i>	<i>Page</i>
2.5.2	Discriminant Analysis Method.	2-11
2.5.3	The Free-Wilson Method	2-11
2.5.4	Substructure Analysis Method (SAM)	2-12
2.5.5	Pattern Recognition Methods	2-12
2.5.6	Overview of Alternative Methods for Data Analysis.	2-13
2.6	Prototype Toxicity Data Bank	2-13
2.6.1	Use of CIDS Fragment Keys for Structure Representation	2-14
2.6.2	Inclusion of Partition Coefficient Data (Log P)	2-14
2.6.3	Implementation of the Toxicity Data Base.	2-15
2.7	Analyses of Data in the Toxicity Data Bank (TDB)	2-15
2.7.1	Empirical Approach	2-15
2.7.2	Methodology Employed.	2-19
2.7.2.1	Discussion of this Method	2-21
2.7.2.2	Reliability of the Estimates Made in Table I.	2-23
2.7.3	Study of TDB by Conventional Statistical Methods	2-24
2.7.3.1	Methods Applied and Results Obtained	2-25
2.7.3.2	Discriminant Analysis	2-25
2.7.3.3	Clustering, Followed by Regression Analysis	2-26
2.7.3.4	Separating the Data Into Three Classes.	2-27

CONTENTS (CONT.)

<i>Section</i>	<i>Title</i>	<i>Page</i>
2.8	Discussion of the Results Obtained from Studies of the Prototype Toxicity Data Bank. . . .	2-28
2.8.1	Discriminant Analysis	2-28
2.8.2	Substructural Analysis Method	2-30
3	RECOMMENDATIONS FOR FUTURE DEVELOPMENT OF THIS PROTO- TYPE DATA BASE	3-1
3.1	Further Expansion of TDB Using Rat Oral LD ₅₀ Data	3-1
3.2	Expansion of the TDB to Include Other Types of Toxicity Data	3-2
3.3	Improvement in Operational Characteristics By Increased Computerization of the TDB	3-3
3.4	Refinement of the CIDS Keys by Experiences Gained From TDB Studies	3-4
3.5	Proposed Use of the TDB by OTS for Early Warning	3-4

APPENDIX A - PHYSICAL PROPERTIES

APPENDIX B - SYSTEM DESIGNED FOR THE TOXICITY DATA BANK

APPENDIX C - DELIVERABLES UNDER CONTRACT

APPENDIX D - REGRESSION EQUATIONS FOR CLASS 1 AND CLASS 2

1. OBJECTIVES OF CONTRACT

1.1 BACKGROUND

By the end of 1973, OTS staff had identified a significant body of scientific literature concerning structure-activity correlations and methodologies. At the same time, no clear-cut decision could be easily obtained as to the potential which these methods might have for helping OTS in its early warning activities. In June 1974, efforts were initiated to review and evaluate this field of scientific endeavor and to develop useful correlations and methodologies for application in the OTS early warning programs.

That this new area was ready for an intensive study by OTS was shown by the scheduling of the first Gordon Research Conference on Quantitative Structure-Activity Relationships in Biology in the summer of 1975.

2. WORK PERFORMED

2.1 LITERATURE SEARCH

Immediately after start-up of the contract, the Office of Toxic Substances furnished 275 full-text hard copies of literature articles dealing with structure-activity correlation which had been located in a preliminary literature search.

The 275 papers served as a nucleus to which was added approximately 75 papers from the personal files of the Principal Investigator. Some 50 additional papers were added over the first four months of the contract as a result of a current-awareness screening operation.

This enlarged bibliography was indexed in depth by SIS literature chemists, and the index was organized by a computerized three-level indexing program. The resulting first version of the index was placed into the NTIS for wider dissemination as NTIS report No. PB 240-658.

Current-awareness screening of the scientific literature received by the FI Library continued for an additional 10 months, during which time about 100 pertinent articles were retrieved. These were augmented by some 50 articles located by SIS personnel assigned to this project. These provided more complete coverage in the area of pattern recognition and discriminant analysis, areas not well covered in the prior retrospective search.

2.2 STUDY OF AVAILABLE TOXICITY DATA

The application of quantitative correlation methodologies requires that the activity or effect be expressed by either continuous or graded incremental quantitative numerical values. Additionally these values should be reproducible at least to $\pm 50\%$ upon repeated measurements, which

is an acceptable amount of variability encountered in making biological activity measurements. Unfortunately, many biological data in the literature are reported without any confidence intervals or other indication of the confidence which one can give to these data.

Data which have been most often successfully studied by the quantitative SA methods include pesticides and pharmacological test data. These test data are frequently reported for more than 10 or 20 closely related structural analogs, and hence represent closely knit sets of data. Classical statistical methods such as regression analysis have been readily applied to these coherent sets of data.

The kinds of biological data which must be considered by OTS for its early warning mission most often represent small units with widely divergent chemical types, and toxicity data is often more descriptive in nature than quantitative.

2.2.1 Toxicity Data in Registry of Toxic Effects of Chemical Substances (RTECS)

The largest collection of toxicity data in the world is reported in the annual Registry of Toxic Effects of Chemical Substances which is prepared by NIOSH. It contains toxicity data for more than 16,000 chemicals, selected from the scientific journal and report literature. Previous editions were called the Toxic Substances List (TSL).

A study of several hundred entries chosen at random from the 1974 TSL showed that more compounds had oral LD₅₀ values in the rat than any other type of toxicity test. The oral mouse LD₅₀ data are second to the rat, and other types of toxicity are much less frequently reported. Although it is recognized that many variable factors enter into an LD₅₀ value, and that such data are not to be considered as "firm" data, the abundance of data suggested the use of the rat oral LD₅₀ as having the best chance for showing structure-activity relationships across a large cross-section of organic compounds.

The 1975 version of TSL, renamed the Registry of Toxic Effects of Chemical Substances, contains a compilation of aquatic toxicity data which may prove useful as a mechanism to locate enough aquatic organism toxicity data to permit study for correlations. The original compilations by Hann and coworkers must be reexamined, since the Registry of Toxic Effects has combined toxicity of many aquatic species into one Table.

2.2.2 Fish Muscle Accumulation Studies

The bioconcentration of highly lipophilic substances in living trout has been successfully correlated with the partition coefficient by Dow scientists (Neely et al, 1975). These workers report a linear relationship between the octanol-water partition coefficient and the bioconcentration factor for a series of halogenated organic compounds. Indeed, a simple "rule of thumb" is suggested by their results, namely, that any halogenated organic compound for which log P (octanol-water) exceeds 5 log units may be considered to pose a potential bioconcentration threat to the environment.

These data are excellent, but too few compounds have yet been studied to permit their inclusion into the prototype toxicity data base.

2.2.3 Carcinogenesis Data

Much of the world's carcinogenesis data is summarized in the PHS Monograph 149 Series. After careful study of these Monographs, and discussions with Dr. Sidney Siegel, (Information and Resources Segment, Carcinogenesis Program, Division of Cancer Cause and Prevention, (NCI) it was concluded that, at present, these data are not easily expressed as one numerical value, since factors such as latency play an important role.

Additional complicating factors include unusual species and strain variations, route of administration, frequency of dosage, frequent lack of adequate control data and conflicting literature reports. Thus the use of established methodologies was not indicated.

2.2.4 Mutagenesis Data

Data on mutagenic activity are becoming available as a result of the expanded use of the "Ames" mutagenesis test as a screening tool for carcinogenic potential of chemical substances. This test makes use of a mutated *Salmonella typhimurium* bacterial strain. A new program is now underway at NCI to establish the consistency of the relationship between mutagenesis and carcinogenesis. Because of the substantial NCI effort, it was felt that manipulation trials with these data were not warranted although it may be advisable to add these data in the near future.

2.2.5 Water Pollution Data

Here the problem was found to be the limitation on the number of compounds for which data exist, although some useful data are given in the November, 1975 Supplement, EPA 440/9-75-009, published by The Hazardous Substances Branch, Office of Water Planning and Standards, EPA. The bulk of the data in STORET is mainly monitoring concentration data, not toxicity data.

The Oil and Hazardous Materials Technical Assistance Data System (OHM-TADS) is an automated system which contains a wide variety of data on more than 850 substances. These data were not readily retrievable when this project began, but the file is now being completed and enlarged. This System is a possible source of data for correlation studies.

2.2.6 Air Pollution Data

An analysis by Hickey, et al (1970), showed some correlations between health effects and current levels of air pollutants. The report was studied for possible reinterpretation of the data. A better correlation was found in our study between high levels of pollutants and respiratory toxic effects one year subsequent to the pollutant levels. These data are epidemiological in nature, and do not provide the type of early warning needed. A visit to the EPA-RTP Air Pollution Center (February,

1975) uncovered only epidemiological and monitoring data in the SAROAD system (SAROAD for "Storage and Retrieval of Aerometric Data). Despite the millions of data items in that system, only some 30 different chemicals were included in these results, an insufficient number for applying correlative techniques.

2.2.7 Teratogenesis Data

The best data collection available at present in this field is Shepard's "Catalog of Teratogenic Agents" (1973). Although it lists 649 agents, the data do not lend themselves to easy analysis or to simple quantitation; the results are mostly descriptive. Several parameters are crucial, including the dose level and route, frequency and timing of administration, types of deformations produced and control problems. The current development of a data bank (teratogenic Information Center, ORNL) may help provide sufficient data for analysis in the future.

2.3 STUDY OF AVAILABLE PHYSICAL-CHEMICAL DATA

Many physical-chemical properties have been used in correlation studies with varying degrees of success. Several of these properties are closely related, or co-linear, and this fact can lead to wrong or unsubstantiated conclusions being drawn from apparently successful correlations. Care must be taken to select those parameters which are substantially independent of each other. Each of the major parameters found to be most useful in establishing correlations is discussed separately below.

2.3.1 Partition Coefficients

The work of Hansch (1971) and Neely (1975) has firmly established the preeminent status of partition coefficient data in this field. These papers substantiate the advantages of using the normal-octanol/water system and record many hundreds of computer-stored experimentally observed values. They also discuss the rationale behind calculations of partition data, although these methods are currently undergoing rather drastic im-

provements, based upon the recent work of Rekker and Nys (1974) (reported at the Gordon Research Conferences, summer, 1975). Basically, the structural fragments of a molecule may be assigned additive fractional partition values and the log of the partition coefficient can be estimated by summation of all fragment partition values.

Partition data alone can often give reasonable correlations with biological data, although they are often used in various combinations with other physical parameters.

The computer printout of partition data from the Pomona College Medicinal Chemistry Project was purchased in this contract and delivered to OTS for use by EPA. Subscriptions to the yearly updates of this compilation are available at an annual charge of \$500. Dr. Albert Leo, Director of the Pomona College Project, has served as a consultant on this project. Use of these partition tables by OTS can provide the "rule of thumb" estimates described for bioconcentration in Section 2.2.2. Preliminary studies for computerized calculation of partition data are being considered by NCI; this activity should be monitored by OTS.

2.3.2 Hammett Sigma Constants

There are currently some 22 varieties of Hammett sigma constants and recently these have been combined into F and R constants (F = field and R = resonance). These constants are used to express the polar (or electronic) effects of substituent groups, usually on a benzene ring. Hence they are limited to use with compounds of this type, a limitation which does not hold for partition coefficients. There is ample precedent for selecting the particular sigma constant for use in a given series of compounds, but the value of these constants when used alone in correlation studies is much more restricted than the use of partition coefficient. Sigma constants usually are employed together with partition data. Tables of the various kinds of sigma constants are readily available in the Pomona College Medicinal Chemical Project's compilation.

2.3.3 Taft Steric Constants

These constants are derivatives of Hammett sigma constants which are derived to express the relative steric strains or crowding effects of substituent groups. Steric constant values are available in the literature for most of the common chemical substituent groups. They have been most constructively applied by Hansch and coworkers to studies of enzyme inhibition, and have little or no applicability in this field outside of such studies.

2.3.4 Molar Refractivity Values

These values can be calculated from chemical fragment refractivity values, as reported by Hansch and coworkers (Journal of Medicinal Chemistry 16, 1207-1216 (1973), in a manner analogous to calculating log P values. The best use of these values in correlation studies has proven to be in studies of enzyme inhibition, as for Taft steric constants. Molar refractivity values are usually employed in conjunction with partition data or other parameters.

2.3.5 Other Types of Physical Constants

Of the several types of physical constants so far discussed, only the partition coefficient and molar refractivity provide values for almost any organic compound, regardless of structure, and these can be calculated for many molecules with a reasonable degree of assurance. Although it is desirable to include constants such as the molecular weight, refractive index, melting and boiling points, etc., for study by pattern recognition methods, one should be aware of the careful attempt reported by Russian workers to relate toxicity of 218 compounds to 38 different physical properties studied one at a time. No reasonable correlations were found using LC_{50} , MAC, narcotic data and any single physical constant. See Appendix A for a list of the 38 physical properties studied. (Ljublina and Filov, 1975).

This background suggests strongly the need to include several properties simultaneously, by pattern recognition methods, as well as the desirability of getting down to a more basic approach by representing the chemical structure instead of using physical constants. It should be pointed out that Hansch has often obtained significant correlations using only the molecular weight, which often mirrors the partition coefficient data for closely related structures.

2.4 METHODS FOR REPRESENTING CHEMICAL STRUCTURES

Chemical structure fragment codes of many types are well-described in the chemical literature. Most of these codes would be applicable to the purposes of this project, but they all are manually encoded, and this is a laborious task which can give rise to occasional errors in coding. The use of augmented connectivity fragments, computer-generated from CAS connectivity tables can overcome the inconsistencies in coding (Lynch, et al, 1971). The augmented connectivity fragment describes every atom by its immediate environment; i.e., by each atom to which it is connected.

However, the limitations of the augmented connectivity fragments are severe, since they do not provide for interrelating structural features more than three atoms apart. One solution of this problem was used by Chu, et al (1975) in which they employed a combination of ring descriptors and chain fragments which connect heterocyclic atoms, together with the augmented atom fragments. However, the chain (or "heteropath") fragments required a large amount of manual editing and correcting, since the existing computer programs were not adequate to make these as exact assignments.

For the developmental purposes of this project, the chemical compounds and their related molecular structures were known; the need was for computer-derived, unambiguous codes for the many substructure fragments contained in each compound.

A readily available source is the U.S. Army Chemical Information Data System (CIDS) which is in operation at Edgewood Arsenal. This system has functioning computer programs to assign fragment code keys based upon a connectivity table representation of each molecule represented in the system. Many of the keys relate to groups of 4, 5 or more atoms, rather than only 3 or 4. Furthermore, the bulk of the chemicals in the Toxic Substances List is available in the CIDS unclassified file. A careful study of the CIDS structure keys was made by Dr. Craig, who decided that these included the basic types of structural relationships which should permit a valid test of the concepts of relating chemical structures to toxicity. An additional point recommending the use of CIDS is that modifications or additions to the keys can be made with relative ease, and the MIDS-EPA group employs the CIDS programs in its chemical substructure search system.

2.5 EVALUATION OF METHODS FOR STRUCTURE-ACTIVITY CORRELATION

2.5.1 Multiple Parameter Regression Analysis (Hansch Method)

Since much information is available on this method, we began our study with a careful evaluation of the requirements for its application to collections of chemical and biological data.

In brief, the Hansch method employs a standard statistical method (multiple regression analysis) to seek significant relationships between a dependent variable (the biological activity) and from one to four or more independent variables, which may be used alone or in combinations. These dependent variables include the physical properties discussed in Section 2.3 above. Experience has shown the partition coefficient to be the most useful single property. The statistical requirements include a minimum number of five compounds for each independent variable studied in an analysis; acceptable correlation coefficients which can range from .7 to .95 or so; and acceptable standard deviations, which rarely are lower than .25 or .3 log units for biological test data.

In summary, the Hansch method is not readily applicable to the wide variety of structures for which toxicity data are available, but can often be applied to smaller sub-sets of closely related chemicals after these have been identified by other techniques. Such uses will represent a "fine-tuning;" of interest, but of little use to OTS for early warning.

The major limitation is that rarely does one have a sufficiently large number of analogs and corresponding biological test data available to permit use of the Hansch method for a new chemical (or even for an existing chemical) proposed for widespread use.

There will occur some specific instances where the Hansch method should prove to be useful to EPA. For example, in the pesticides field the submission of a close structure analog to a series of well-studied products should allow the predictive application of a multiple parameter regression equation derived by the Hansch method for the well-known series of compounds, permitting an educated guess as to the relative hazard posed by the new compound. However, it is the current accepted practice to require a battery of biological tests before a new pesticide is permitted to be marketed. The mere knowledge that a new substance has a close structural relationship to known toxic pesticides serves as the most readily applied "Early Warning" indicator, whether or not a quantitative estimate of the toxicity potential has been made.

The Hansch computer program for multiple parameter regression analysis was purchased from the Pomona College Medicinal Chemistry Project for use by OTS and EPA; the computerized data bank of sigma constants, octanol-water and other partition data was also obtained and delivered to OTS. Thus OTS now has the required computer programs and partition and sigma constants listings for using the Hansch method, when the proper data are available for its application.

2.5.2 Discriminant Analysis Method

Since the more rigorous requirements of the Hansch method preclude its use in most cases of interest to OTS, the less demanding statistical method, discriminant analysis, was analyzed and evaluated for its potential application to the needs of OTS.

The basic requirements and limitations of this method were studied, and the required computer programs were shown to be readily available and easily used at any major computer facility. Basically the several methods for applying discriminate analysis employ statistical methods to discriminate between two or more groups of substances by deriving an optimal line, plane or multidimensional plane which will permit useful discrimination between these groups and can be used to predict whether a new substance has a high probability of belonging to one or another group, e.g. carcinogenic, non-carcinogenic, etc.

The prototype data base, containing rat oral LD₅₀ data, was studied successfully by this method, as discussed in Section 2.7.3. Discriminant Analysis thus offers a readily applicable method for identifying potentially toxic substances, and can often serve the desired goals of OTS by quickly classifying a new substance as having a "high toxicity" or "low toxicity".

2.5.3 The Free-Wilson Method

This is a classical method which has been well-described and contrasted with the Hansch method (Craig, 1973).

The basic assumption of this method is the constancy (and additivity) played by substituent groups in a closely related set of structural analogs. This assumption is tested by the solution of a series of multiple equations with multiple unknowns. This method of solution imposes certain rigorous mathematical requirements such that a solution cannot be obtained without considerable data. Unfortunately for the need for OTS, one rarely will encounter a sufficient number of closely

related, systematically designed analogs to permit use of the Free-Wilson method, save in the design and study of drugs or pesticides. No applications of this method to a diverse group of compounds is known.

2.5.4 Substructure Analysis Method (SAM)

This method shares with the Free-Wilson method the basic assumption that a structural unit of a molecule makes an additive and consistent contribution towards the overall biological activity of each molecule which contains that structural unit, but unlike the Free-Wilson method it is not limited to use with close structural analogues.

This method was first described in 1974 (Cramer, et al 1974). It was designed to overcome the limitations of the Free-Wilson method, and thus permits the use of any organic chemical regardless of any particular structural relationships. To accomplish this a very large training set of many hundreds of structures is required. The method employed by Cramer, et al, involved the assignment of a substructure activity frequency to each of the various fragments of a molecule (obtained from a manually assigned chemical fragment coding system) and summing up the mean substructure activity frequencies of all the substructures present in the compound under consideration. The resulting calculated "Mean Substructure Activity Frequency" for a compound then was found to distinguish significantly between active and inactive compounds in a crude biological screening test.

2.5.5 Pattern Recognition Methods

Although there are many types of pattern recognition methods known, two of the methods most often studied in this field will be discussed here.

The "nearest neighbor" method involves computing Euclidean distances between points in n-dimensional space, and by use of a training set of compounds, a new substance is classified into the particular group for which the sum of the squares of the Euclidean distances to all group members is smallest for the new substance.

The "learning machine" method is an iterative error-correcting procedure which creates a hyperplane in n-dimensional space which can discriminate between the known "active" and "inactive" compounds in a training set.

These two methods are well illustrated in a paper by Chu et al (1975). Limitations of these methods are also well discussed therein.

In the preliminary studies of a prototype toxicity data base (Section 2.7.3.3) the use of a pattern recognition method did not give any better results than could be accomplished by the much simpler method of discriminant analysis. Although we recommend keeping in close touch with developments in pattern recognition, the relative success of the far more easily applied methods of substructure analysis and discriminant analysis in the studies of the prototype data base make it unnecessary to go over to the more complex pattern recognition methods at this time for the purposes of OTS.

2.5.6 Overview of Alternative Methods for Data Analysis

A study was made of the many types of mathematical methods available for organizing, analyzing and evaluating kinds of data, to see whether one or several existing methods could be readily adapted to the needs of OTS. The study included consideration of the properties of the data elements (e.g., nominal, ordinal interval or ratio) and of the mathematical tools available for each type of data. It soon became evident that empirical methods should be considered when a number of diverse compounds are being studied.

2.6 PROTOTYPE TOXICITY DATA BANK

The analyses of methods and data types indicated that the best test of utility would be obtained from establishing a prototype toxicity data bank and testing various methods on the actual data on hand. From the reports of multiple regression methods, (including the Hansch, Free-Wilson and discriminant analysis approaches) it was apparent that em-

pirical methods for studying the data should be first considered, with possible application of more refined statistical methods when suitable sets of data are identified.

In designing the data base it was concluded that at least 500 compounds would be required to achieve a critical size to permit valid studies of the type envisaged. Since the toxicity data were the critical factor, it was decided to choose compounds on a roughly random basis, from those for which an oral rat LD₅₀ value was reported in the 1974 TSL. [The 1975 successor (RTECS) to the 1974 TSL was not yet available when this phase was initiated.]

Some 800 compounds were selected from the sections A through H of the TSL. This set was considered adequate to serve for the prototype study.

2.6.1 Use of CIDS Fragment Keys for Structure Representation

Chemical structural representation was provided by the CIDS keys which vary from 4 to 8 or more characters in length. Although they were computer-assigned to all compounds in the CIDS file, the system was not designed to display individual keys assigned to each compound. Therefore special programs were written to permit the extraction and print-out of all keys assigned to each compound requested.

2.6.2 Inclusion of Partition Coefficient Data (Log P)

Fewer than 20% of the more than 800 compounds selected for the TDB were reported in the partition data banks of Professor Hansch. Dr. Albert J. Leo, Pomona College, a consultant to this project, calculated log P values for compounds for which no experimental log P values were known, with the exception of a few structures for which insufficient knowledge exists to permit the calculation to be made. These values were desired in order to permit inclusion of partition data in discriminant analysis, regression analysis and pattern recognition studies.

2.6.3 Implementation of the Toxicity Data Base (TDB)

Of the more than 800 compounds selected for the TDB, 700 were found to exist in the Edgewood Arsenal non-classified file. CIDS keys for 687 compounds, were extracted for the TDB. The required programs and sub-routines were developed under a subcontract.

Under another subcontract the necessary computer programs were prepared to merge the CIDS keys with the toxicity, molecular formula and log P values into the master TDB record. A brief description of the system designed for the TDB is given in Appendix B.

2.7 ANALYSES OF DATA IN THE TOXICITY DATA BANK (TDB)

2.7.1 Empirical Approach

The TDB master record was searched by computer to obtain the following printouts:

- a) A listing ordered by CIDS keys.

This list contains the Army Registry Number (ARN) (a compound identifier for the CIDS system) and oral rat LD₅₀ values, ordered in ascending value of toxicity, for each compound to which that key was assigned. Since the toxicities range in value from 1.11 to 6.88 (log 1/C values), this permits a ready count of the distribution between six categories of log 1/C for each key. These categories are:

log 1/C = 1.0 to 1.999	= Category 1
2.0 to 2.999	= Category 2
3.0 to 3.999	= Category 3
4.0 to 4.999	= Category 4
5.0 to 5.999	= Category 5
6.0 to 6.999	= Category 6

b) A listing ordered by rat oral LD₅₀ log 1/C values.

This list gives for each compound the ARN, mouse oral LD₅₀ 1/C value, each CIDS structure key, molecular weight and log P value (n-octanol-water). The distribution of compounds in each of the six categories is obtained from this list.

c) An "explosion" run of more than 6,000 pages listed for every key assigned to an individual compound, the complete records for all other compounds in the data base to which that key was also assigned. This provides a valuable resource to study combinations of keys in search of more discriminatory power for correlations.

Our modifications of the method reported by Cramer, et. al., include the use of quantal grades for toxicity and the use of frequency of assignment for each key across quantal grades to estimate the activity of the molecule as a composite described by combination of these keys. This method has been applied manually to estimate toxicities for 31 compounds including 10 not in the original TDB. Results are summarized in Table I.

Table I

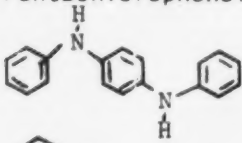
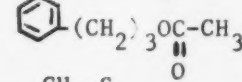
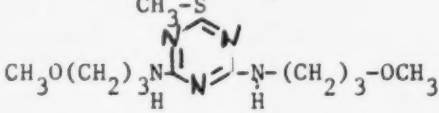
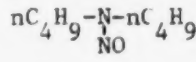
		Estimated Log 1/C	Observed Log 1/C	Deviation
1.	Caffeine	3.82	3.02	.80
2.	Pentachlorophenol	3.30	3.99	-.69
3.		3.10	2.04	1.06
4.		2.43	1.58	.85
5.		2.74	2.33	.41
6.		2.44	2.12	.32

Table I (continued)

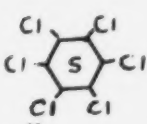
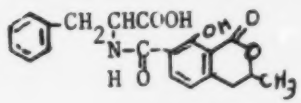
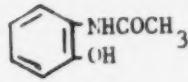
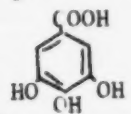
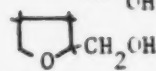
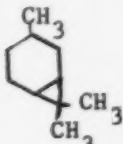
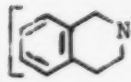
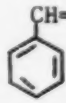
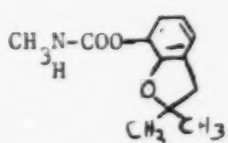
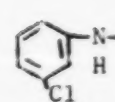
		Estimated Log 1/C	Observed Log 1/C	Deviation
7.	$\text{CH}_3-\text{N}(\text{NO})-\text{nC}_4\text{H}_9$	2.52	2.95	-.43
8.	$\text{C}_2\text{H}_5-\text{N}(\text{NO})-\text{C}_2\text{H}_5$	3.02	2.67	.35
9.	 Lindane	3.46	3.52	-.06
10.	$\text{H}_2\text{N}-\text{CH}_2\text{CH}_2-\text{N}(\text{H})-\text{CH}_2\text{CH}_2-\text{N}(\text{H})-\text{CH}_2\text{CH}_2\text{OH}$	2.49	1.50	.99
11.	CH_3NHNH_2 *	3.26	3.24	.02
12.	Alloxan **	3.97	1.50	2.47
13.		3.84	4.30	-.46
14.		2.96	1.47	1.49
15.	$\text{CH}_2=\text{CH}-\text{OCH}=\text{O}$	2.16	1.41	.75
16.	$\text{CH}_3\text{O}-\text{C}(=\text{O})-\text{CH}=\text{C}(\text{O}-\text{P}(\text{OCH}_3)_2)-\text{CH}_2\text{COOCH}_3$ *	2.52	3.96	-1.44
17.	DDT	3.12	3.49	-.37
18.	Cl_3CCOOH	2.59	1.58	1.01
19.		3.27	1.53	1.74
20.		2.10	1.83	.27
21.	FCH_2COONa *	3.03	5.65	-2.62

Table I (continued)

		Estimated Log 1/C	Observed Log 1/C	Deviation
22.	$(\text{CH}_3\text{O})_2\text{-P(=O)(OH)-CHCCl}_3$ *	2.52	1.54	.98
23.		2.09	1.44	.65
24.	 $\cdot \text{H}_2\text{SO}_4$ *	2.79	3.70	-.91
25.		2.45	1.32	1.13
26.	$(\text{C}_2\text{H}_5\text{O})_2\text{-P(=S)-S-CH}_2\text{CON-(C}_2\text{H}_5)_2$ *	2.97	4.34	-1.37
27.	$\text{CH}_3\text{COCH}_2\text{CON-(C}_2\text{H}_5)_2$	3.00	1.51	1.49
28.	$\text{Cl-CH}_2\text{CH}_2\text{-OH}$	2.48	3.05	-.57
29.	$\text{nC}_6\text{H}_{13}\text{-OH}$	2.36	1.34	1.02
30.		4.20	4.44	-.24
31.		2.93	2.25	.68

* Insufficient number of examples of one key in TDB to permit confidence in this value.

**Unusual toxicity, no close structural analog present in TDB.

2.7.2 Methodology Employed

To make the estimations reported in Table I the following empirical steps were manually employed to calculate the activity for compound 31, Table I.

- a) The chemical to be considered is encoded by assignment of the CIDS structure keys, using the *Handbook of CIDS Chemical Search Keys* if the compound is not in the CIDS system.
- b) Locate the frequency of encoding for each assigned key from the frequency listing. Do not use keys assigned fewer than 5 times if possible, since the use of keys assigned to fewer than 5 compounds introduces a considerable possibility for error. One can use these rarely assigned keys when necessary but caution should be observed in interpreting the results. Keys assigned to more than 400 compounds have but little chance of being significantly related to toxicity in this TDB, which contains a very small number of highly toxic substances.
- c) If no SCN key exists (Specific Cyclic Nucleus) for a cyclic moiety, use all GCN (Generic Cyclic Nucleus) keys. If an SCN key is found, do not use the GCN keys, to avoid redundancy.
- d) Obtain the distribution counts for the number of compounds in each of the 6 toxicity classes for each key, using the listing of rat log 1/C values ordered by keys (see
- e) Prepare a substructure vector matrix for each key as follows:

	<u>Toxicity Range:</u>					
Key:	1	2	3	4	5	6
FG51R	0	4	11	3	0	0

(this means that no compounds in the TDB, assigned the FG51R key, had toxicities ranging from log 1/C = 1.1 to 1.99; 4 ranged from log 1/C = 2.0 to 2.99; 11 ranged from 3.0 to 3.99, etc.). The last compound in Table I will serve as the example for the following steps.

f) Sum all 6 substructure vectors as follows:

	<u>Toxicity Ranges:</u>					
Key:	1	2	3	4	5	6
FG51R	0	4	11	3	0	0
DACN=2	36	62	9	1	1	0
HR11E	4	17	4	1	0	0
FG268R	28	129	48	15	6	0
SCN48	99	209	38	7	1	1
Sum of Vectors	167	421	110	27	8	1

g) Since the total data base (687 compounds) was not evenly divided between classes 1, 2, 3, 4, 5 and 6, the weighting factors of 3, 2, 6, 12, 60 and 120 were used to correct for the unequal distributions in the data bases. This is necessary to prevent the great preponderance of high frequency data from drowning out the important low frequency data, which, however, carries the really important information concerning toxicity. If the usual method for obtaining averages were applied without this weighting step (a "normalizing" operation), the skewed nature of the data would prevent successful calculations. The origin of the weighting factors arises from the following distribution which was found for this TDB: The total tallies for all 681 compounds by toxicity levels are:

	TOXIC LEVEL 1	TOXIC LEVEL 2	TOXIC LEVEL 3	TOXIC LEVEL 4	TOXIC LEVEL 5	TOXIC LEVEL 6	TOXIC LEVEL MISSING
POPULATION OF COMPOUNDS ≈	200	300	100	50	10	5	26

Ignoring the 26 missing data, the 665 compounds remaining have toxicity level tallies which are all fractions of 600; the weighting factors of 3, 2, 6, 12, 60 and 120 result from dividing 600 by each toxicity level tally.

Sum of Vectors =	167	421	110	27	8	1
Correction Factor =	<u>x 3</u>	<u>x 2</u>	<u>x 6</u>	<u>x 12</u>	<u>x 60</u>	<u>x 120</u>
Normalized Vector Sum =	501	842	660	324	480	120

Total sum of these six vector sums is 2,927.

- h) To obtain weighted averages, the normalized vectors are multiplied by 1, 2, 3, 4, 5, and 6, and the sum of these vectors, divided by the normalized vector sum is the estimated toxicity value:

501	842	660	324	480	120
x 1	x 2	x 3	x 4	x 5	x 6
501	1684	1980	1296	2400	720

Sum of these 6 values is 8,581; $\frac{8,581}{2,927} = 2.93 = \text{estimate of toxicity.}$

- i) This estimated value (2.93) is compared with the value reported in the 1974 TSL (compound FD 80500) of 1200 mg/kg. Since the molecular weight is 213.7, $\frac{1.2g}{213.7} = .005615$ moles/kg; to convert this to log 1/C it is necessary to take the reciprocal ($= \frac{1}{.005615}$) = 178.1;

the log value for which is 2.25. This is then to be compared with the estimated value of 2.93; a deviation of 0.68 is noted. This deviation is in the "safe" direction of predicting a somewhat greater toxicity than that actually observed.

2.7.2.1 Discussion of this Method

The rat log 1/C values are grouped into six "quantal" units. Since these are log units, .5 corresponds to a 3-fold difference in the actual value for an LD₅₀.

When the entire range from 1.1 to 6.8 (a 5.7 log or 500,000-fold range) is considered, the ability to estimate rat oral toxicity values to even ± 1 unit should be useful.

It must be emphasized that the technique presented is applicable to any range desired, such as 3 or 4 quantal units instead of 6; also to any type of activity for which data can be located or measured across a large set of compounds, and for which the basic assumption of additivity is found to hold.

Among the interesting conclusions which can be arrived at by a quick scan of the frequency counts by keys for each toxicity range is the relative "toxicity" for a given key.

This can be readily estimated as follows:

Key SCN48 (the benzene ring):

	1	2	3	4	5	6
	99	209	38	7	1	1
	<u>x3</u>	<u>x2</u>	<u>x6</u>	<u>x12</u>	<u>x60</u>	<u>x120</u>
(Sum = 1207)	297	418	228	84	60	120
	<u>x1</u>	<u>x2</u>	<u>x3</u>	<u>x4</u>	<u>x5</u>	<u>x6</u>
(Sum = 3173)	297	836	684	336	300	720

$\frac{3173}{1207} = 2.63$; this represents the "average" toxicity due to a benzene ring.

Thus, the "average" toxicity values for the structure keys are determined by the sum of all the experiences included in the TDB. Obviously the larger the TDB, the more reliable will be the estimates of toxicity, both for the fragment keys and for the entire molecule calculations. With the total of 687 compounds employed the principle is well-illustrated by the examples in Table I, but this TDB currently lacks an adequate number of examples of several important functional groups to permit their inclusion in toxicity estimates.

The concept of a dynamic data base is important to the successful implementation of the SAM. The computerized nature of TDB permits ready expansion and updating of the data. Equally important is the flexibility designed in TDB to permit facile addition of other types of toxicity data, such as aquatic organism toxicity, mouse toxicity, carcinogenicity, mutagenicity, or any other type of biological data which can be measured and graded. In each case, separate printouts would be prepared similar

to that described for the rat LD₅₀ data. In an expanded TDB system it would be practical to maintain all of the files on-line or in a minicomputer; the required distributions could then be retrieved as needed or printed out from time to time to maintain updated listings.

2.7.2.2 Reliability of the Estimates Made in Table I

The estimated rat log 1/C values for each compound listed in Table I are plotted against the observed rat log 1/C values (taken from the 1974 Toxic Substances List) in Figure 2-1. Included in this plot are compounds (marked by asterisks) for which too few examples of related compounds exist at present in TDB to permit confidence in the estimates (fewer than 5 examples of important FG or SCN keys). They are included to show that estimates can be made even in these cases, but the user is cautioned to be aware of the lack of confidence in that estimate.

The FG and SCN keys which have fewer than 5 examples in the TDB, and which are coded for those compounds in Fig. 2-2 marked with an asterisk, represent the following structural features:

hydrazine - NHNH_2

dialkylphosphate $\text{RO}-\text{P}(\text{O})_2-\text{OR}$

guanidino $\text{-NH}-\text{C}(\text{NH}_2)=\text{NH}$

hexahydropyrimidine 

dithiophosphate $\text{-rS}-\text{P}(\text{O})_2-\text{OR}$

The addition of more examples containing these structural features to the TDB will improve the utility of the TDB.

The remaining 23 compounds in Fig. 2-1 without asterisks all were estimated to be at least one-tenth as toxic as the observed values, while 6 of the 23 were estimated to be more than 10 times as toxic as observed. Seventeen of the 23 compounds were estimated to be within ± 1 log unit (from one-tenth to ten times the actual toxicities. When the full range of more than 500,000 between the least toxic and most toxic compounds in the TDB is considered, the results are quite reasonable.

It must be emphasized at this point that the present TDB is too small to permit its predictive use with confidence except for relatively simple chemical structures which contain only those functional groups and ring structures which are well-represented in the prototype TDB. As is true of all systems which generate information based upon a "training set" (in this case, 687 compounds), it is misleading to use uniquely (or rarely) occurring groups to predict the activities of the compound in which they exist. However, 9/10 of the compounds in Table I which do not exist in the TDB gave satisfactory estimates ranging from +.02 to +1.06 log units higher than the observed values. The one compound poorly predicted, alloxan, was estimated to be almost 300 times more toxic than the reported LD_{50} . Aside from the point already made that too few close analogs of alloxan exist in the TDB at present, it is interesting to observe that alloxan is known to be highly toxic to the rat upon chronic administration, and it is used to produce a state in rats akin to diabetes in man. Although TDB is presently based on lethal oral toxicity data in the rat, it is interesting to speculate that the accumulated information in TDB already is suggestive of "toxicity" in a broader sense than the input which was used.

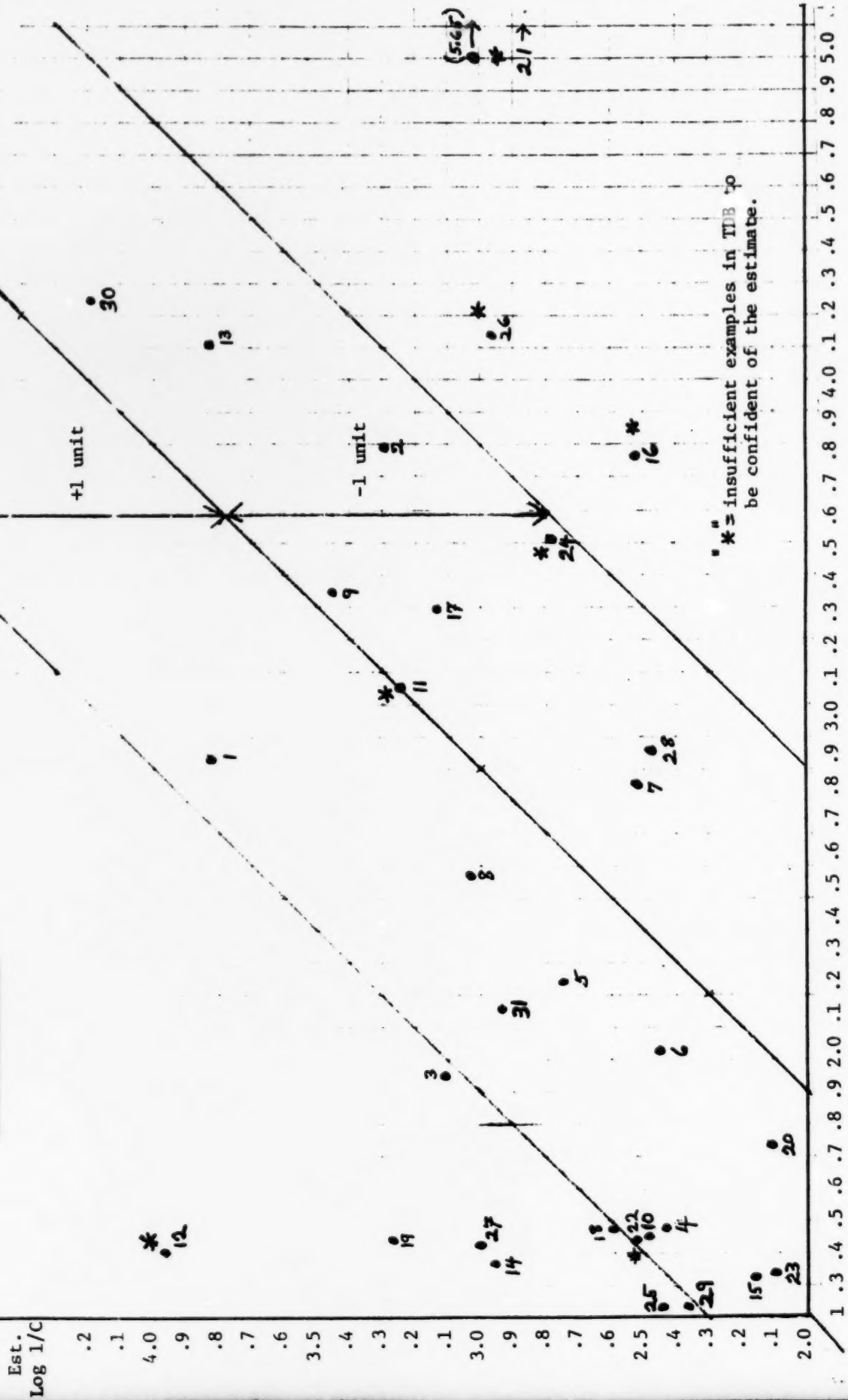
2.7.3 Study of TDB by Conventional Statistical Methods

Under a subcontract all of the data contained in the prototype TDB was analyzed by multivariate techniques, including discriminant analysis, regression analysis and preliminary clustering techniques. The data

Figure 2-1

Plot of Observed vs Estimated $\log_{10}$ Values

For Oral Rat LD 50



set consisted of the 549 compounds for which rat toxicity, log P (partition data), molecular formulas and the CIDS fragment keys were all available.

2.7.3.1 Methods Applied and Results Obtained

Initially only the 101 fragment keys which had been assigned 10 or more times were studied. A regression equation containing 30 variables was obtained for 543 compounds by stepwise-regression. The standard error of estimate was .54 and R^2 was .33 ($R=.574$). All 101 keys had an opportunity to be entered into the regression. It is interesting to note the standard error of .54, which supports our earlier decision to assume a range of ± 1 log unit for the estimates discussed in Section 2.7.2 above. This range is approximately twice the standard error found this preliminary regression study.

2.7.3.2 Discriminant Analysis

A series of stepwise discriminant analyses was performed, based upon the upper and lower quartiles of log 1/C values; a typical result is shown below:

<u>Actual Toxicity</u>	<u>Classified as:</u>		<u>Error Rate %</u>
	<u>High</u>	<u>Low</u>	
High	91	33	27
Low	12	132	8

(overall average error rate = 17%)

It should be noted that the skewed nature of the data results in a range of 1.114 to 1.968 for the lower quartile, and a much broader range of 2.727 to 6.914 for the higher quartile. The extremes of the two quartiles are only 0.759 log 1/C units apart. Thus the relatively poor discrimination found is not surprising.

2.7.3.3 Clustering, Followed by Regression Analysis

Next, clusters were sought, using the ISOGEN method, which is a modified learning machine method. A random set of 142 compounds was clustered, using 137 variables, after scaling all variables to lie between 0 and 1. $\log P$, $(\log P)^2$, molecular weight and 134 CIDS keys were used. One main cluster of 129 compounds, together with 3 clusters of 6, 2 and 4, respectively, were obtained.

It was found that 15 CIDS keys did not discriminate among these four clusters, so by eliminating these keys a larger group of 305 randomly selected compounds could be used. This resulted in 2 clusters, of 297 and 7 compounds.

Limiting the regression study to this large cluster, they then obtained the following results:

Run #2: 41 variables, 297 compounds:

$$R^2 = .63$$

Standard Error of Estimate = .5396

Standard Deviation of Residuals = .4999

Major variables: $(\log P)^2$, molecular weight; FG120;

FG51R; FG80; FG96; SCN48; SCN49; GCN3=C2,01;

GCN3=C5; GCN4=C5,01; GCB4=C6; GCN6=1,2; GCN1=3

GCN5=1

Run #3: 30 variables, 285 compounds:

$$R^2 = .50$$

S.E. of Est. = .4619

S.D. of Res. = .4368

Major variables: molecular weight; NCN=3, FG120; FG51R;

FG80; FG94; FG96; HR12R; GCN= C8

Run #4: 38 variables, 267 compounds:

$$R^2 = .56$$

S.E. of Est. = .3625

S.D. of Res. = .3356

Major variables: molecular weight; DACN=2; FG120; FG51R;

FG80; FG96; HR12R; GCN5=0; GCN5=6.

The difference between runs 2, 3 and 4 is that certain outliers were not included in runs 3 and 4, thus sharpening the resulting equations. These preliminary results pinpoint certain keys to contain significant importance as contributors to oral rat toxicity.

2.7.3.4 Separating the Data Into Three Classes

The compounds were then divided into three classes:

- a) Class I: those compounds in which no key occurs with a frequency of use of less than 7.
- b) Class II: those compounds containing at least one rare key.
- c) Class III: the outliers, mostly those compounds with high toxicity.

The intent was to separate class III from classes I and II by discriminant analysis and then derive separate regression equations for classes I and II. In actual practice this TDB contains most of the highly toxic substances in class III.

This approach was followed through, and led to the suggestion that a three-step system be used for classification and prediction of the rat toxicity of chemicals:

First, the discriminant equations would be used to pinpoint the class III compounds (highly toxic), thus achieving a highly desirable early warning result;

Second, the remaining compounds would be separated into classes I and II, based upon the presence or absence of rare keys;

Third, toxicity for members of these two groups would be predicted using the appropriate regression equations. It should be noted that this approach can handle the actual log 1/C values as a continuous function, whereas substructural analysis is more easily applied to step-wise or quantal gradations in toxicity.

The above approach was carried out to give regression equations for group I and II compounds and a discriminant equation to separate group III from groups I and II. Details of the regression equations are given in Appendix D.

The discriminant analysis equation needs to be adjusted to assure that no false negatives slip by. This adjustment can be readily made; it will result in an increase (not prohibitive) in the false positives.

One of the major advantages of the discriminant analysis approach over substructural analysis is that one does not need a training set (or large data base) to apply discriminant analysis, whereas the substructural analysis approach, in common with pattern recognition methods, requires a large training set to establish baselines.

2.8 DISCUSSION OF THE RESULTS OBTAINED FROM STUDIES OF THE PROTOTYPE TOXICITY DATA BANK

The studies of the prototype TDB by two quite different methods for structure-activity correlation have shown that both methods have the potential to alert OTS about potential toxicity for novel or untested compounds.

2.8.1 Discriminant Analysis

This method provides a model equation which discriminates between highly toxic substances for the rat and the bulk of the non-toxic substances, and which is based upon only the presence of one or more of a group of 12 structural keys. Although this group of structural keys may require some additions as new data are added, it is worthwhile to examine these at this point.

<u>Variable</u>	<u>Class (I+II)</u>	<u>Class III</u>	<u>F</u>
FG205	1.24	43.3	82.1
GCN3=C5	.983	18.9	65.6
FG29R	1.24	43.3	41.2
SCN107	.095	38.4	33.6
SCN84	.095	38.4	33.6
SCN45	.727	13.1	30.8
GCN4=C3	1.24	22.3	20.5
HRIR	1.15	4.90	15.9
FG118	.852	13.5	11.0
FG34	.852	13.5	11.0
FG112	1.16	5.30	10.8
FG231R	-.488	14.7	10.2
Constants	-.140	-13.1	

Class I contains those compounds for which all the structure keys are used at least seven times in the TDB. Class II contains all compounds with at least one rare key (assigned less than 7 times). Class III contains the outliers, mostly those compounds with high toxicity. The above table describes the Discriminant Equations and shows the different coefficients for the keys in Class III compounds and Classes I and II compounds (combined). The F test indicates the significance of the differences between these coefficients; e.g., between 1.24 and 43.3 for FG205. (An F value above 2 or 4 is usually significant at a 95% level, hence all of these are highly significant.) It should be realized that the discriminant approach merely differentiates between classes, e.g., toxic and non-toxic and should not be used to estimate relative toxicities.

Below are given the structural meanings of these keys:

<u>KEY</u>	<u>STRUCTURE</u>	
FG205	phosphorodithioate	
GCN3=C5	cyclopentyl ring	
FG29R	imide group	C=N — (group attached to ring)
SCN107	isochroman	
SCN84	2,3-dihydrobenzofuran	
SCN45	piperidine	
G-CN4=C3	cyclopropyl	
HR1R	ring-methyl	
FG118	acetylenic	
FG34	primary amide	
FG112	aliphatic halogen	
FG231R	trialkyl-phosphate	

These structural groupings range from highly specific to very generic. However, they alone can serve as alerting tools to OTS for purposes of early warning. They are not sufficient by themselves to pinpoint a highly toxic substance, but they can be legitimately used to question a new structure, as far as potential rat toxicity is concerned.

Each new set of toxicity data entered into the TDB should be studied by discriminant analysis to see if a similar set of discriminant equations can be derived from that set. The resulting keys can provide a quick rule of thumb for questioning further the potential toxicity of a novel substance which contains keys permitting discrimination between high and low toxicity.

In the particular example just studied (rat oral toxicity) the Class I and Class II compounds were also studied by regression analyses, so that in the event that a toxicity arises from a particular combination of relatively non-toxic and common keys, the regression equations may be applied to predict toxicity. These equations are given in Appendix D. As more compounds are added, these equations can readily be derived anew.

2.8.2 Substructural Analysis Method

This relatively simple method has been shown in this project to lead to quite acceptable predictions of toxicity. This approach can be applied to all types of toxicity data which can be obtained (or located) in graded or quantitative terms of activity. Although these initial results have been obtained by manual use of computer-generated print-outs, it should be emphasized that estimates will be more readily obtained from computerization of the calculations, since these are often laborious when done manually.

The same point should be made concerning the manual assignment of CIDS keys; these are normally assigned by computer, can lead to inaccurate encoding by an inexperienced chemist. Of interest is the type of feedback which these studies can provide for increasing the fragment key specificity to represent toxicological significance.

Because the development of the CIDS system was for generic retrieval of chemical structures, it is understandable that the keys which are chosen for that purpose will not always be the best keys for relating structures to biological activity. Using this study of the TDB to guide improvements in the CIDS keys is a logical step, which would be required regardless of what fragment encoding system were to be used.

The recent interest in the CIDS systems shown both by NCI and EPA scientists and computer personnel could serve to expedite any future work in this area. It is especially noteworthy that current programming is already underway to provide for assignment of CIDS fragment keys by computer programs designed to convert the Chemical Abstracts Services connectivity tables into the CIDS keys. This will have the result of making readily available the CIDS keys for over 3,500,000 compounds already on record at CAS. The ability to access the CIDS keys by computer would make the use of this empirical method readily operable by OTS.

The results obtained to date from this study support the validity of the basic assumption, namely; that structural fragments can be assigned incremental toxicity values which may serve to estimate the resulting toxicity of the entire molecule. This result was by no means a foregone conclusion, especially when the complex nature of lethality and the problem of poor duplicability of LD₅₀ results are considered. The size of the TDB was crucial; a large enough number of compounds had to be studied to override the variability factors. It is unlikely that similar results could have been obtained with confidence using a much smaller data base.

Although the TDB should be increased in size before being widely applied, it can even now be used to estimate toxicity (oral rat LD₅₀) for those chemicals which do not contain CIDS structural keys which occur fewer than five times in the TDB. Many hundreds of chemicals of interest to OTS fall into this category, and could be easily calculated if the SAM method were further computerized.

3. RECOMMENDATIONS FOR FUTURE DEVELOPMENT OF THIS PROTOTYPE DATA BASE

Possible ways to further develop the dual approach to using SAR methods for purposes of early warning include:

- a) Further expansion of the TDB using rat oral LD₅₀ data.
- b) Expansion of the TDB to include other types of toxicity data, such as marine toxicity data, carcinogenicity, etc.
- c) Improvement in operational characteristics by increased computerization of TDB.
- d) Refinement and additions to CIDS structure keys, by experience gained in studied of the TDB.

Each of these points is discussed below.

3.1 FURTHER EXPANSION OF TDB USING RAT ORAL LD₅₀ DATA

The best way to increase the size of the rat oral LD₅₀ data in the TDB is to extract all such data from the NIOSH computer tapes which are now available. By including the CAS registry numbers (also on the NIOSH tapes) it should then be possible to obtain the CIDS structure keys by computer from the connectivity tables. Towards this end, MDSD-EPA has already made arrangements with CAS to obtain connection tables by supplying them with the CAS registry numbers.

Should an unexpected delay be encountered in these arrangements, it is recommended that the remainder of the 1974 TSL (I through Z) be scanned manually for inclusion of more compounds with oral rat toxicity data, as we did in setting up this prototype TDB.

One problem which is frequently encountered is the change in rat oral LD₅₀ values reported in the 1975 TSL as compared to the 1974

edition. Some of these changes are so startling as to suggest that one or the other must be reported in error (e.g., a change from the microgram to milligram level). These apparent discrepancies should be checked by examination of the original literature references, to avoid entering possible noise into the TDB. It would be desirable, but time-consuming, to check all reported values; NIOSH is supposed to be correcting such errors and may be able to provide the desired confirmation.

3.2 EXPANSION OF THE TDB TO INCLUDE OTHER TYPES OF TOXICITY DATA

It should be emphasized that inclusion of the chemical structure keys for compounds without any accompanying toxicity data is not encouraged for this stage of development of TDB. At some time in the future it may be desirable to do so to provide a "shopping list" for compounds available for toxicity study to increase the number of examples of rare structure groups (or keys). However, at this stage it makes much more sense to add other toxicity data, both for compounds already in the TDB and for new structures.

The Carcinogenicity problem is of great concern to EPA. Although the problems of determining relative degree of carcinogenicity are staggering, and NCI is struggling with this great problem, it would appear that sufficient data now exist to permit tentative assignment of several quantal grades of carcinogenicity, and the following are suggested for consideration:

Highly active carcinogens (aflatoxins, 2-AAF, etc.). This would be used for potent carcinogens shown to cause cancer consistently in several species.

Active carcinogens (vinyl chloride, hexamethylpyrophosphoramide, nitrosamines, etc.) Some of these may belong in the first category.

Suspected carcinogens (conflicting results, only one species, currently under trial.)

Not carcinogenic (reserved for compounds adequately tested).

Compiling a test set of approximately 200 to 300 compounds with these categories into the TDB, and generating the CIDS frequently counts for the chemical structure keys would be the first step. Ultimately an estimate of the carcinogenic potential would be obtained.

As more mutagenicity data becomes available, it should be possible to treat it in a manner similar to the carcinogenicity data. Indeed, it should be easier to quantitate the results from the Ames test than from a carcinogenesis bioassay. Further impetus for this approach may come from the careful study underway at present at NCI to confirm or deny the proposed relationship between carcinogenicity and mutagenicity.

From a practical point of view, the present TDB contains much data for the mouse oral LD₅₀ test. These data have not been studied in the way reported for the rat, but could readily be examined.

3.3 IMPROVEMENT IN OPERATIONAL CHARACTERISTICS BY INCREASED COMPUTERIZATION OF THE TDB

There are several ways to improve the accuracy and ease of operation of the prototype TDB. With but minimal programming, one can readily calculate the average toxicity associated with all of the keys, as shown by a manual calculation for the benzene ring (Section 2.5.2.1). This should be done to provide estimates of those keys most associated with toxicity, to see whether any surprises are encountered.

With additional programming, the TDB calculations shown in Section 2.5.2 can be systematized so that the process will not only be easier and faster, but it should be freed from possible arithmetic errors. With but little additional effort, the calculation process can be placed into an on-line computer mode. The whole TDB operation can also be converted to a minicomputer operation, which may offer several advantages. During these programming stages the system can still be studied in the manual mode.

3.4 REFINEMENT OF THE CIDS KEYS BY EXPERIENCES GAINED FROM TDB STUDIES

The current CIDS keys were assigned by organic chemists after careful study of the problems in retrieval of chemical structures which are chemically related to each other. The requirements for this type of "generic" retrieval system are not necessarily the same as the requirements for relating chemical structure to biological activity. Even with the TDB at its present level of 687 compounds, our preliminary experiences have pinpointed several keys for which it is advisable to have greater specificity of structural information.

It will become increasingly obvious with more experience that this type of feedback should result eventually in the best type of keys for this purpose; namely, keys designed to optimize the biological differences between chemical structure fragments.

3.5 USE OF THE TDB

Even in this prototype state, the TDB can be used as described in Section 2.7.3 and 2.7.2 to predict toxicities for many compounds in the rat. When aquatic organism toxicity data, carcinogenesis data and possibly mutagenesis data are added, predictions of several types of toxicity may be possible for many types of chemicals, and each new set of toxicity data will strengthen the utility of this data base. We recommend that both the discriminant analysis and substructural analysis methods be employed at present since we have not had sufficient opportunity to compare these methods and to determine whether one or the other, or both, should always be employed.

REFERENCES

1. Adamson, G. W., Lynch, M. F., and Town, W. G., *J. Chemical Society C*, 1971, 3702-3706.
2. Chu, K. C., Feldmann, R. J., Shapiro, M. B., Hazard, Jr., G. F., and Geran, R. I., *J. Medicinal Chemistry* 18, No. 6, 539-545 (1975).
3. Craig, P. N., "Advances in Chemistry, Volume 114, Chapter 8, 1973, pp. 115-129, *American Chemical Society*.
4. Cramer III, R. D., Redel, G., Berkoff, C. L., *J. Medicinal Chemistry* 17 533 (1974).
5. Hickey, R. J., Boyce, D. E., Harner E. B. and Clelland, R. C., *IEEE Trans. on Geoscience Electronics*, October, 1979, pp. 186-201.
6. Leo, A. J., Hansch, C. H., and Elkins, D., *Chemical Reviews* 71, 525-616 (1971).
7. Ljublina, E. I. and Filov, V. A., "Methods Used in the USSR for Establishing Biologically Safe Levels of Toxic Substances", World Health Organization, Geneva, 1975, Chapter 2, pp. 19-44.
8. Neeley, W. B., Bronson, D. R., and Blau, G. E., *Environmental Science and Technology* 8, 1113-1115 (1974).
9. Nys, G. G. and Rekker, R. F., *Chim. Ther.* 8, 521 (1974).
10. Shepard, T. H., "Catalog of Teratogenic Agents", The Johns Hopkins University Press, 1973.

APPENDIX A

PHYSICAL PROPERTIES

studied by Ljublina and Filov, USSR

- | | |
|---|---|
| 1. Molecular weight | 22. Logarithm of distribution coefficient (water/air) |
| 2. Density | 23. Surface tension |
| 3. Molar volume | 24. Kinematic viscosity |
| 4. Refractive index | 25. Dynamic viscosity |
| 5. Molar refraction | 26. Solubility |
| 6. Melting point | 27. Specific heat capacity |
| 7. Boiling point | 28. Specific heat of vapor |
| 8. Saturated vapor pressure | 29. Thermal conductivity |
| 9. Saturated vapor pressure | 30. Atomic polarization |
| 10. Equilibrium temperature | 31. Parachor |
| 11. Rate of change of t_{boil} with pressure | 32. Electric dipole moment |
| 12. Rate of change of t_{boil} with pressure | 33. Dielectric constant |
| 13. Critical density | 34. Specific dispersion |
| 14. Critical temperature | 35. Absolute dispersion |
| 15. Critical pressure | 36. Primary ionization potential |
| 16. Latent heat of fusion | 37. Entropy of liquid |
| 17. Latent heat of vaporization | 38. Entropy of gas |
| 18. Heat of combustion | |
| 19. Heat of formation of gas | |
| 20. Helmholtz energy of formation of gas | |
| 21. Logarithm of distribution coefficient (olive oil/water) | |

APPENDIX B

SYSTEM DESIGNED FOR THE TOXICITY DATA BANK

A flexible system was designed to permit the resulting TDB to expand to many thousands of compounds, should this be desirable.

Although the system is not limited by hardware considerations, a brief description of the system designed for this particular application is given. Provision was made for a record length of 1316 characters, unblocked, and the system was mounted on an IBM 360-40 with 384K bytes of main-frame memory, 4 tape-drives and a 2314 disk unit. Nine-track tape at 1600 bpi and EBCDIC codes were employed. The batch system can readily be upgraded to an on-line system, should this become desirable. The programs are written in ANSI COBOL under OS: an assembly language call program was written to invoke a Fortran log function to permit use of logarithms in calculations of $\log \frac{1}{C}$ values.

The programs permit the accession of all compounds having any desired combination of parameters in common. The parameters include the toxicity data, structural keys, and physical constants such as log P values, but there is provision for many other parameters if desired.

APPENDIX C

DELIVERABLES UNDER CONTRACT

A. Reports

1. Index to Subjects and Authors for Structure-Activity Correlation Bibliography: 68 pages. Available from NTIS, PB 240-658.
2. The Use of the Hansch Multiple Parameter Method of Structure-Activity Correlation to Identify the Hazard Potential of Environmentally Significant Chemicals: Paul N. Craig & Jon E. Villaume, 11 pages, 1 Nov. 1974.
3. The Use of Classification to predict the Biological Activity of Environmentally Significant Chemicals: Discriminant Analysis: Jon E. Villaume, 14 pages, 17 March 1975.
4. Review of paper by R. J. Hickey et. al.; 'Ecological statistical studies concerning environmental pollution and chronic disease': J. H. Waite, 22 pages, August 1975.
5. Overview of Alternate Methods for Data Analysis, by John Waite, 75 pp., April 1976.
6. 'Models for Biochemical Toxicity', by Genessee Computer Center, Inc.: 30 pages, 19 February 1976. (Performed under subcontract to FIRL.)
7. A Computer Program to Extract and Display CIDS Keys, by Fein-Marquart Associates, Inc: 208 pages, December 1975. (Performed under subcontract to FIRL.)

B. Data Files

1. Partition coefficient Data Bank from the Pomona College Medicinal Chemistry Project. (Approximately 500 pages of computer print-out.) Includes Sigma Constant Data Bank.
2. Computer Program (Hansch-3) for Multiple Parameter Regression Analysis. (About 1,000 punched cards.)
3. Computer printouts from Toxicity Data Bank:
 - a) A listing ordered by CIDS keys which contains Army Registry numbers and oral rat LD50 data in ascending order of toxicity for each key (from Cryptanalytic)

- b) A listing ordered by rat oral LD₅₀ values (as log 1/C (from Cryptanalytic)).
- c) A listing of compounds for which each key was assigned, ordered by key (from Fein-Marquart).

C. Computer Programs Written for this Project

1. Frequency distributions of the Fein-Marquart structure key data (24 byte octal numbers) and print programs.
2. Extraction of applicable structure keys from the mass of Fein-Marquart data.
3. Conversion of the 24 byte octal keys into the conventional CIDS keys.
4. Frequency distributions of C-3 above, and associated print programs.
5. Generation of the sample toxic substances data base, which included:
 - a. Merging (3 data sources) of Fein-Marquart data with FIRL-supplied data (CAS number, rat and mouse LD₅₀, Log P, Molecular formulae).
 - b. Error detection and correction for above.
 - c. Sort sample data base in ARN order.
 - d. Print sample data base.
6. The 'explosion' program, including molecular weight and Log 1/C calculations and many associated computer runs, prints, etc.
7. A computer program to order Rat Log (Log 1/C) within substructure keys and its associated sort, sum, and print programs.

APPENDIX D

REGRESSION EQUATIONS FOR CLASS I AND CLASS II COMPOUNDS

A. Regression Equation for Class I Compounds (See Section 2.7.3.3)

"Best" Compromise Class I Regression Equation

<u>Variable</u>	<u>Coefficient</u>	<u>Std. Error</u>	<u>F</u>
FG112	.783+00	.116	45.9
MW	.302-02	.000	45.4
FG120	.318+00	.093	11.7
HR1R	.336+00	.102	10.9
NCN=0	.143+01	.483	8.73
FG83	.212+00	.087	5.95
GCN5=3	-.257+00	.107	5.80
FG268R	.198+00	.084	5.53
FG96	-.214+00	.094	5.16
A-C=0	.109+01	.501	4.77
GCN4=C2N	-.742+00	.455	2.66
(Log P) ²	-.878-02	.005	2.60
FG96R	-.209+00	.149	1.96
FG144	.162+00	.122	1.76
Constant	.188+01		

$$R^2 = .6277$$

$$R^2 = .394$$

$$S.E. = .442$$

$$S.D. = .432$$

$$N.V. = 14$$

B. Regression Equation for Class II Compounds (See Section 2.7.3.3)

"Best" Compromise Class II Regression Equation

<u>Variable</u>	<u>Coefficient</u>	<u>Std. Error</u>	<u>F</u>
MW	.217-02	.001	22.5
FG80	.617+00	.151	16.8
Log P	-.782-01	.019	16.5
FG154R	.618+00	.185	11.1
FG51R	.634+00	.190	11.1
FG120	.430+00	.135	10.2
GCN2=3	.524+00	.198	7.00
GCN4=C5N1	.647+00	.247	6.85
GCN5=6	.415+00	.173	5.77
HR1E	.219+00	.092	5.68
FG34R	-.545+00	.230	5.64
DACN=6	.471+00	.202	5.46
GCN5=5	.598+00	.289	4.27
FG120R	-.341+00	.189	3.24
GCN3=C5	.463+00	.258	3.22
FG24R	.346+00	.226	2.35
HR2ER	-.216+00	.156	1.92

(Continued)

APPENDIX D (Continued)

<u>Variable</u>	<u>Coefficient</u>	<u>Std. Error</u>	<u>F</u>
GCN2=5	-.236+00	.171	1.91
GCN4=C10	-.401+00	.291	1.91
GCN3=C501	.235+00	.208	1.89
FG178R	-.209+00	.157	1.77
Constant	.205+01		

R = .6922

R² = .479

S.E. = .490

S.D. = .466

N.V. = 21

Both of these regression equations were derived by stepwise regression analysis, allowing new variables to be entered as long as they made significant contributions to the reduction of the residual variance.